



D8.2

Data Management Plan

31 July 2023

Work package:	8 - Project Management	
Author(s):	POLYTECHNEIO KRITIS	
Reviewer #1	SML	Leticia Pascual
Reviewer #2	NKUA	Dimitris Gizopoulos
Dissemination Level	PU	
Nature		

This document may contain proprietary material of certain REBECCA contractors. The commercial use of any information contained in this document may require a license from the proprietor of that information.

Date	Author(s)	Comments	Version	Status
19/07/2023	K. Georgopoulos/I. Mavroidis	1 st Draft	v0.1	
21/07/2023	D. Gizopoulos	Review #1	v0.1	
25/07/2023	L. Pascual	Review #2	v0.1	
27/07/2023	K. Georgopoulos	Addressed Reviewers' Comments	v0.2	



Executive Summary

This deliverable describes the Data Management Plan (DMP) of the REBECCA project, and the approach the partners will follow on open source and open data. It also describes the corresponding data management procedures to be followed for the use cases of the project.



Table of Contents

Executive Summary	1
Table of Contents	2
List of Figures	3
List of Tables	3
1. Introduction	4
1.1 Reference Documents and Methodology	4
2. Data Management Plan	5
2.1 Technical Datasets	5
2.2 Scientific Publications and Reports	8
2.3 Evaluation Datasets	9
2.4 Other Datasets and other Information/Data	9
2.5 FAIR Data	10
2.5.1 Making Data Findable	10
2.5.2 Making Data Openly Accessible	10
2.5.3 Making Data Interoperable	11
2.5.4 Increase Data Reuse	11
3. Data Protection	12
4. References	12



List of Figures

Figure 1: Screenshot of REBECCA website showing links to major sub-menus	6
--	---

List of Tables

Table 1: Up-to-Date Technical Datasets	6
Table 2: REBECCA Existing Press Releases & Dissemination Actions	8
Table 3: List of REBECCA deliverables in relation to their dissemination level	8
Table 4: Estimated Evaluation Datasets for the REBECCA project	9
Table 5: Other materials expected to be generated within REBECCA	9



1. Introduction

REBECCA is a collaborative project, featuring a large consortium of partners and a highly distributed research team working on it. In order to have a common understanding it is important to determine the data management and exchange procedures and guidelines. This document contains all the up-to-date related information for the proper functioning of the project, and it will be updated when needed during the project lifetime.

1.1 Reference Documents and Methodology

The preparation of this document has been based on the following project documents:

- REBECCA Grant Agreement (GA-101097224) and its annexes, and specifically the Description of Action (DoA).
- REBECCA Consortium Agreement.



2. Data Management Plan

This Data Management Plan (DMP) describes the data management life cycle for all datasets that will be collected, processed or generated by the REBECCA project. This document outlines how research data will be handled during the project and after the project completion.

The DMP is not a fixed document, but it evolves along with the REBECCA project. This is the first version of the DMP. Afterwards, it will be updated on a yearly basis according to the project's needs.

The following categories of datasets are identified within the REBECCA project:

- Technical datasets including the technical work, such as source code of tools, libraries, etc.
- Scientific publications that describe the research work within REBECCA.
- Evaluation datasets that accompany the scientific publications and usually provide more information than that included in the publications.
- Other material, such as the Consortium Agreement, Grant Agreement, leaflet, ethical forms, consent forms, documentations, or any other kind of dissemination material.

The following Sections provide important information for each of the aforementioned categories.

2.1 Technical Datasets

In short, the technical datasets include the code of the REBECCA platform and its applications. Most of the technical datasets will be maintained at the project's Github repository in order to be accessible by all the REBECCA partners. In particular, the binary and/or the source code of each technical dataset will be maintained as decided by the partner that owns it.

Moreover, it is likely that some of the technical datasets will be provided under Open Source licenses and, therefore, they will be made publicly available on the REBECCA website and Github, under a suitable sub-menu that facilitates the source code download of the technical datasets. A typical screenshot of the REBECCA website is shown in Figure 1. The Open Source code will be available for an additional period after the end of the REBECCA project, most likely in the order of two years beyond the completion of the work.

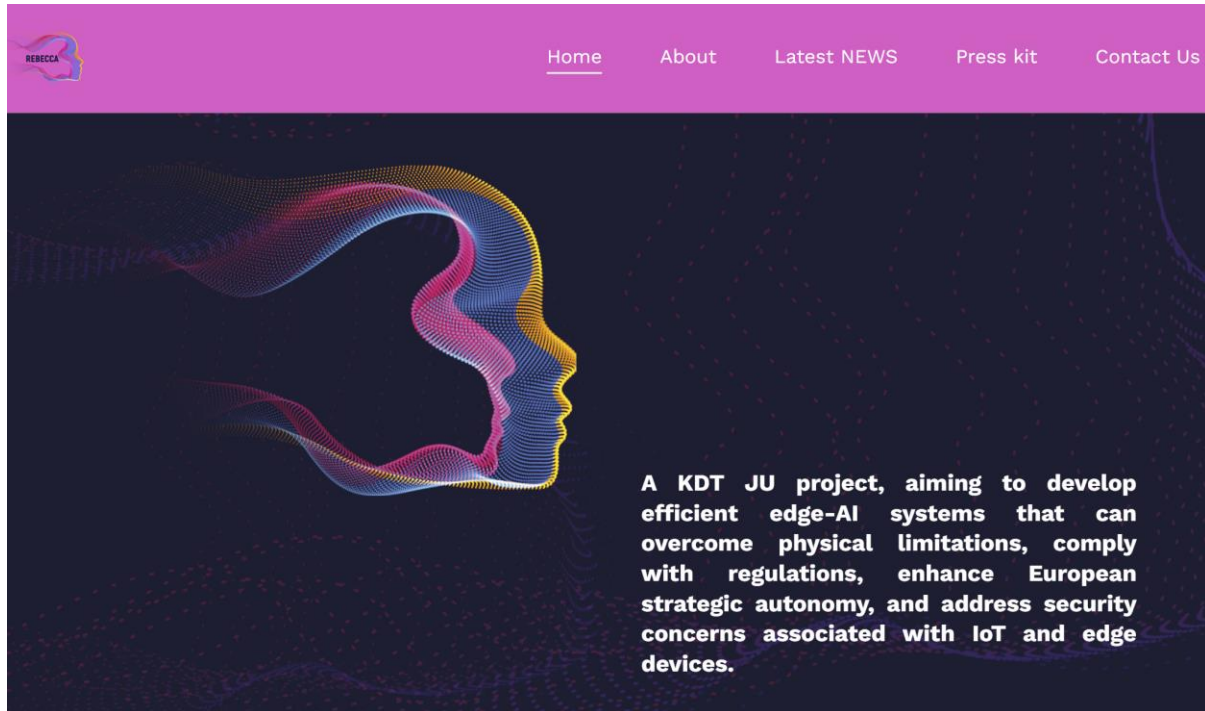


Figure 1: Screenshot of REBECCA website showing links to major sub-menus

The foreseen technical data, which will be updated throughout the project, are the following.

Table 1: Up-to-Date Technical Datasets

Type	Dataset Name	Description	Sharing	Archiving and Preservation
Design Files and netlists	REBECCA_RISCV	RISC-V processing system	Github for internal sharing, REBECCA website for public access.	Dataset will be available at least two years after the end of the project.
Design Files and netlists	REBECCA_Interconnect	Interconnect	Github for internal sharing, REBECCA website for public access.	Dataset will be available at least two years after the end of the project.
Design Files and netlists	REBECCA_Chip2Chip	Chip2Chip communication	Github for internal sharing, REBECCA website for public access.	Dataset will be available at least two years after the end of the project.
Source code	REBECCA_Neuromorphic	Neuromorphic AI Accelerator	Github for internal sharing, REBECCA website for public access.	Dataset will be available at least two years after the end of the project.
Source code	REBECCA_ML	Hierarchical ML accelerator	Github for internal sharing, REBECCA website for public access.	Dataset will be available at least two years after the end of the project.
Source code	REBECCA_CNN	CNN Accelerator	Github for internal sharing, REBECCA website for public access.	Dataset will be available at least two years after the end of the project.
Source code	REBECCA_IDS	Intrusion Detection and Prevention System	Github for internal sharing, REBECCA website for public access.	Dataset will be available at least two years after the end of the project.

Source code of tools	REBECCA_RISCV	Chiplet Integration Methods and Tools	Github on https://github.com/REBECCA-CHIP for internal or shared access	Dataset will be available at least two years after the end of the project.
SW Libraries	REBECCA_Hypervisor	RISC-V optimized Hypervisors for AI accelerators	Github on https://github.com/REBECCA-CHIP for internal or shared access	Dataset will be available at least two years after the end of the project.
SW Libraries	REBECCA_AILib	AI Libraries optimized for RISC-V, AI accelerators and/or Reconfigurable logic	Github on https://github.com/REBECCA-CHIP for internal or shared access	Dataset will be available at least two years after the end of the project.
SW Libraries	REBECCA_EDDL	EDDL Library optimized for RISC-V and AI accelerators	Github on https://github.com/REBECCA-CHIP for internal or shared access	Dataset will be available at least two years after the end of the project.
Source code of tools	REBECCA_Design Space	Design Space Exploration toolset	Github on https://github.com/REBECCA-CHIP for internal or shared access	Dataset will be available at least two years after the end of the project.
Source code	REBECCA_VSLAM	Implementation of a very precise VSLAM model based on inspection input.	Github for internal or shared access	Dataset will be available at least two years after the end of the project.
Source code	REBECCA_Filters	Implementation of filters for fuzzy and over complete pictures.	Github for internal or shared access	Dataset will be available at least two years after the end of the project.
Source code	REBECCA_Meshroom	Meshroom extensions to speed up the generation of 3D inspection models.	Github for internal or shared access	Dataset will be available at least two years after the end of the project.
Source code	REBECCA_UAV	Inference AI-based module for defect detection directly on-board the UAV	Github for internal or shared access	Dataset will be available at least two years after the end of the project.
Source code	REBECCA_AIAsset	AI Asset workflow for end-to-end deployment & benchmarking, allowing full integration and validation	Github for internal or shared access	Dataset will be available at least two years after the end of the project.
Source code	REBECCA_DamageInspection	High-speed AI-based damage inspection and dynamic event-based vision-in-the-loop for semiconductor	Github for internal or shared access	Dataset will be available at least two years after the end of the project.

		manufacturing equipment		
--	--	-------------------------	--	--

2.2 Scientific Publications and Reports

This Section keeps track of the project publications and reports. Since the project is in its early stages, the table below includes one press release and potential future dissemination actions.

Table 2: REBECCA Existing Press Releases & Dissemination Actions

Type	Dataset Name	Description	Sharing	Archiving and Preservation
Press Release	REBECCA_PR_Spanish	Spanish Press Release	Google Drive for internal or shared access	Data will be available at least two years after the end of the project
Edge AI Conference publication	REBECCA_Edge AI	Edge AI Conference to be organized by KDT projects	REBECCA website and OpenAIRE	

The list of publications will be amended throughout the project in order to include new publications during the project lifetime. The publications and related research data will be publicly provided at research data repositories, respecting the policies and rules set out by the publishers (journals or conferences). In particular, the Open Access Infrastructure for Research in Europe [1] will be the main point for providing public access to the research publications of REBECCA. Moreover, several of them will be posted on the REBECCA website as well as on the open data repositories of individual partners' websites.

Apart from the above publications, REBECCA's deliverables are considered to be scientific reports of the work within the project. In addition, several deliverables will be publicly available and posted on the REBECCA website. Overall, the REBECCA technical deliverable outcomes are as follows.

Table 3: List of REBECCA deliverables in relation to their dissemination level

Type	Deliverable	Dissemination Level	Sharing	Archiving and Preservation
Document, Report, Demonstrator	D1.1, D1.2, D1.3, D1.5, D1.6, D2.1, D2.2, D2.3, D2.4, D5.1, D5.2, D6.1, D6.3, D7.3, D7.4, D7.6, D7.7, D7.9, D7.10, D8.1, D8.5, D8.6, D8.7, D9.1, D9.2, D9.3, D9.4	Sensitive	Google Drive for internal sharing, EC repository ¹	Data will be available at least two years after the end of the project.
Document, Report	D1.4, D3.1, D3.2, D3.3, D3.4, D4.1, D4.2, D4.3, D4.4, D6.2, D6.4, D7.1, D7.2, D7.5, D7.8, D8.2, D8.3, D8.4,	Public	Google Drive for internal sharing, EC repository	Data will be available at least two years after the end of the project.

¹ EC repository provided for REBECCA which can be accessed by the members of the consortium and the EC

2.3 Evaluation Datasets

The evaluation datasets accompany the scientific publications and reports. These datasets include evaluation measurements of the REBECCA platform, such as performance, throughput, latency, energy consumption and frames per second. Moreover, these datasets include videos and photos that might be captured during the evaluation of the REBECCA platform.

We foresee the following evaluation datasets and believe that any likely changes will not deviate significantly from what we present in Table 2.

Table 4: Estimated Evaluation Datasets for the REBECCA project

Type	Dataset Name	Description	Sharing	Archiving and Preservation
Measurements	REBECCA_IoT_benchmarks_eval	Profile measurements of REBECCA IoT devices using common benchmarks.	Github for internal sharing, OpenAIRE for public access.	Data will be available at least two years after the end of the project.
Measurements	REBECCA_UC1_eval, REBECCA_UC2_eval, REBECCA_UC3_eval REBECCA_UC4_eval	Profile measurements of the REBECCA Use Cases on the REBECCA infrastructure.	Github for internal sharing, OpenAIRE for public access.	Data will be available at least two years after the end of the project.

2.4 Other Datasets and other Information/Data

Apart from the aforementioned data, several other documents will be produced within REBECCA, such as the Consortium Agreement, Grant Agreement, leaflet, ethical forms, consent forms, documentations, minutes and presentations at conference calls, brochure, leaflet, and website files. A list of these data is provided in Table 5.

Table 5: Other materials expected to be generated within REBECCA

Type	Dataset Name	Description	Sharing	Archiving and Preservation
Document	REBECCA_GA	GA signed by all partners in electronic format.	Google Drive for internal sharing, EC repository	Data on VCS will be available at least two years after the end of the project.
Document	REBECCA_CA	Consortium Agreement	All partners have the same hard copy of the CA. An electronic copy is provided on Google Drive.	Hard copy must be maintained and preserved by each partner even after the end of the project.
Document	REBECCA_ConfCall_X	Minutes and presentations of monthly conference calls.	Google Drive for internal sharing.	Data will be available at least two years after the end of the project.
Files	REBECCA_website	REBECCA website	Teraglobus is responsible for the	Data will be available at least

			content of the webpage	two years after the end of the project.
--	--	--	------------------------	---

Along the project lifecycle, more dissemination material than the one presented in Table 3 will be produced, such as presentations, videos, and others. This material is planned to be made available through the website and other dissemination channels, e.g. https://twitter.com/kdt_rebecca and <https://www.linkedin.com/company/rebecca-kdt-ju/>. The partners will consider the legislation and policies at the European level in order to protect data, including some particularities from different countries.

2.5 FAIR Data

The REBECCA project follows the “FAIR Guiding Principles” for scientific data management, for improving the (F)indability, (A)ccessibility, (I)nteroperability and (R)euse (FAIR) of digital assets. These principles emphasize on the capacity of computational systems to find, access, interoperate and reuse data with none or minimal human intervention.

2.5.1 Making Data Findable

The REBECCA Technical data will be organized by application and system configuration, making the results easy to navigate and find. The results will be in a standard format based on the output of the applications and, thus, they will be relatively small for any standard identification mechanisms.

The input data for the applications, will be either auto-generated or from publicly available sources when possible, and there is no plan of collecting sensitive personal information of any kind. Also, there are no plans to persist the data, other than the method used to generate or obtain the input data. Therefore, no unique identifiers (i.e., Digital Object Identifiers) are required to locate those data.

2.5.2 Making Data Openly Accessible

Any source code developed in the REBECCA project will be made available when possible, and acceptable by the owning partner. Open source repositories will be used, such as GitHub, with a standard structure and relevant documentation. Any changes made to the source files of upstream projects will be produced as PR (Pull Request) or a patch to an existing repository, so as to be distributed as open-source in the corresponding project or their community.

All input and output data generated in this project may follow the Open Access policy. However, when there is a huge amount of data used as part of the application dataset, only the generation and/or source of the datasets for the applications may be provided and become accessible by the community through the repository to demonstrate the validity of the implementation of our hardware, software or system design.

The source code for any components licensed as open source will be included in a public Github repository, and it will be linked to repositories of the relevant software tools (e.g. Linux) when applicable. When needed, new Git projects and repositories will be created for the various parts of the REBECCA project. Furthermore, Git submodules will be used to link the integrated version with the corresponding Git project of each of the components (software and/or hardware and/or system), when applicable.



For any models, code, or other information developed by either REBECCA project partners, or third parties, an access agreement with the owner may be required in some cases.

2.5.3 Making Data Interoperable

The datasets that will be used to evaluate the performance of REBECCA will not rely on a specific data format; they will consist of smaller fragments (of larger applications) that will be used to demonstrate the advances in the chip/system. The data generated when the REBECCA systems will be running the project's reference applications, will compare the performance with that reported by other sources, to better determine the advances on the programming and runtime support. This information will be included in scientific documents to properly determine the advances of the architectures used in the REBECCA project.

The evaluation results (traces, benchmark output, etc.) will be published with enough detail in order to allow other scientists to compare their results with the ones generated in this project, as well as research communities.

2.5.4 Increase Data Reuse

During the project, data reuse options for the datasets will be dealt with on a case by case basis, striving to keep them public, accessible, free of charge and reusable under request. A standard reuse license, such as the Creative Commons license will be used. This type of license-use is important for industrial exploitation, following the obligations set out in the Grant Agreement (GA). For cases where there are limitations, either due to inherited licenses or business practices, on a case by case basis we will try to find a solution to make the data reusable and shareable. The goal is to enable the widest reuse or inheriting the license types from the different data sources, and taking care of each partners' business constraints or legal limitations on them.



3. Data Protection

There is the possibility of collecting personal data from EU citizens through surveys or even social networks. In order to comply with the European legislation, the project will follow the principles of Directive 95/46/EC [2] regarding data protection.

This means that all personal data collected, inside or outside the consortium, will be treated following the next principles:

- **Notice:** subjects must know that their data is being collected.
- **Purpose:** data must be used for the stated purpose and no other.
- **Consent:** data cannot be disclosed or shared without explicit consent.
- **Security:** data must be kept safe and secure.
- **Disclosure:** subjects must be informed as well as parties collecting data.
- **Access:** subjects should grant access to their personal data and be allowed to correct it.
- **Accountability:** subjects should be able to hold collectors accountable for adhering to these principles.

Regulator Authorities will be notified when needed before any processing of personal information.

However, there is no plan of collecting sensitive personal information such as bank statements, credit card numbers or addresses, but it is necessary to take this into account even for minor actions such as anonymizing surveys.

4. References

- [1] <https://www.openair.com/>
- [2] <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A31995L0046>